Journal of Innovative Optical Health Sciences
Vol. 7, No. 6 (2014) 1450028 (7 pages)

DOI: 10.1142/S179354581450028X

World Scientific
www.worldscientific.com

Geographical classification of Nanfeng mandarin by near infrared spectroscopy coupled with chemometrics methods

Xuan Zhang*, Yiping Du*,‡, Peijin Tong*, Yuanlong Wei†
and Man Wang*
*Shanghai Key Laboratory of Functional Materials Chemistry
and Research Center of Analysis and Test
East China University of Science and Technology
Meilong Rd 130, Shanghai, P. R. China 200237

†Comprehensive Technology Center of
Jiangxi Entry-Exit Inspection and Quarantine Bureau
and Jiangxi Province Engineering Research Center
of Infrared Spectroscopy Application
South Gan River Avenue 2666, Nanchang
Jiangxi Province, P. R. China 330038
‡yipingdu@ecust.edu.cn

Received 3 September 2013
Accepted 2 January 2014
Published 7 March 2014

Near infrared spectroscopy (NIRS), coupled with principal component analysis and wavelength selection techniques, has been used to develop a robust and reliable reduced-spectrum classification model for determining the geographical origins of Nanfeng mandarins. The application of the changeable size moving window principal component analysis (CSMWPCA) provided a notably improved classification model, with correct classification rates of 92.00%, 100.00%, 90.00%, 100.00%, 100.00%, 100.00% and 100.00% for Fujian, Guangxi, Hunan, Baishe, Baofeng, Qiawan, Sanxi samples, respectively, as well as, a total classification rate of 97.52% in the wavelength range from 1007 to 1296 nm. To test and apply the proposed method, the procedure was applied to the analysis of 59 samples in an independent test set. Good identification results (correct rate of 96.61%) were also received. The improvement achieved by the application of CSMWPCA method was particularly remarkable when taking the low complexities of the final model (290 variables) into account. The results of the study showed the great potential of NIRS as a fast, nondestructive and environmentally acceptable method for the rapid and reliable determination for geographical classification of Nanfeng mandarins.

Keywords: Near-infrared spectroscopy; Nanfeng mandarin; geographical origin; changeable size moving window principal component analysis; variable selection.

1. Introduction

Nanfeng mandarin, native to Nanfeng county, Jiangxi province, as one of the famous and precious citrus varieties in China, is a kind of distinctive products of geographical indication with a very long cultivation history. It is an extremely popular product and an important international traded commodity in Nangfeng county, with its features of thin skin, soft and succulent pulp, sweet and sour tastes, intense in aroma, unique flavor and seedless. Thus, Chinese government has established the corresponding national standard on the protection of this citrus variety.[a] Furthermore, owning to its excellent quality, it has been widely naturalized all over China, e.g., Fujian, Guangxi, Hunan province, etc. The Nanfeng mandarin, introduced to other places in China grows well and shows similar appearance, however, there are differences in taste to some extent. Meanwhile, with the development of international trade and improvement of quality of people's life, the requirement of product quality is higher than before. Therefore, with the aim of guaranteeing authenticity and protecting the consumer from fraudulent labeling of mandarin, a means of differentiating mandarins from different geographical locations must be devised.

In recent years, near infrared spectroscopy (NIRS) with the advanced features of fast, simple, cheap and nondestructive, has attracted considerable attention for the qualitative and quantitative analyses in food and agriculture industry.[1,2] For the purpose of quality control, NIRS data have been effectively combined with multivariate techniques such as cluster analysis (CA),[3] principal component analysis (PCA).[4] and artificial neural network (ANN),[5] etc. to identify authenticity, producing area, as well as similar varieties of the analyzed samples. For instance, there are many domestic and international scholars using NIRS technique to identify different varieties of coffee,[6,7] juicy peach,[8] tea,[9] tobacco,[10] etc. Likewise, many researchers have reported the authentication and geographical classification of honey,[11] olive oils,[12,13] red wines,[14] cheese,[15] apple juice,[16] etc. In our group,[17] the feasibility of identification of Nanfeng mandarins from varied regions employing NIRS and PCA has been studied, and declared better classification results. However, the very small sample size and the absence of independent prediction set of this study influenced the applicability and robustness of the classification models. Moreover, the samples from different villages/towns in Nanfeng county Jiangxi province were not well separated when taking all samples from other provinces into account.

The aim of the present study is precisely to propose a strategy for developing improved and reliable classification model for accurate geographical identification of Nanfeng mandarin, which is essential for assessing mandarin quality, based on their NIR spectra. The PCA was selected as class-modeling method for this study, since it is a powerful data mining technique in multivariate calibration of spectral analysis, and can perform both numerical and graphical results. In this way, changeable size moving window partial least squares (CSMWPLS)[18] algorithm was modified and coupled with PCA to construct a variable selection method called changeable size moving window principal component analysis (CSMWPCA), that was applied on these spectra in order to improve the performance of classification models and reduce the size of datasets in calibration and validation process. For evaluating the effect of the wavelength selection technique on sample classification, the results obtained before and after feature selection were analyzed and compared.

2. Experimental Methods

2.1. *Samples preparation*

A total of 583 mandarin samples were harvested from 7 different geographical origins, which are Fujian, Guangxi, Hunan province, and four different villages/towns in Nanfeng county Jiangxi province, as shown in Table 1. In the trial, in order to

Table 1. The origins of the samples in the research.

Category No.	Origin	Number of samples	
		Calibration set	Validation set
C1	Fujian, China	50	6
C2	Guangxi, China	54	6
C3	Hunan, China	90	10
C4	Baishe, Jiangxi, China	89	10
C5	Baofang, Jiangxi, China	89	10
C6	Qiawan, Jiangxi, China	90	10
C7	Sanxi, Jiangxi, China	62	7

[a]GB19051-2003.

guarantee the representation of mandarin samples and extend the applicability of the classification models, the samples of each category explored in this experiment were chosen in the different orchards of the same area, meanwhile, samples at different heights and sunlight conditions were also considered in every orchard.

2.2. *Recording of NIR spectra*

All NIR spectra of Nanfeng mandarin samples between 1000 and 1800 nm were obtained using a SupNIR-1000 portable near infrared spectrometer (Focused Photonics Inc., Hangzhou, China) in diffuse reflectance mode, equipped with a tungsten halogen lamp light source and an InGaAs detector. A fiber-optics probe diffuse reflectance accessory was placed on the surface of the intact mandarin sample to collect spectra at ambient temperature (ca. 298 K) with 10 scans and a resolution of 1 nm. And in order to reduce the error of operation, the probe was covered with a tin foil paper to keep a beam diameter of 1 cm. Each mandarin was measured three times at three equally equatorial positions and the average spectrum of three parallel measurements was used. Spectra were recorded in random order and a reference spectrum was measured with 15 min interval during the spectra measurement.

2.3. *Theory and algorithm*

2.3.1. *Pretreatment of measured NIR spectra*

The measured NIR spectra always comprise substantial information derived from sample attributes, as well as environmental and instrumental information which strongly take an effect on the performance of the analysis system. In order to remove the scattering effect created by diffuse reflectance, decrease baseline shifts, overlapping peak and the detrimental effects on the signal-to-noise ratio, multiplicative scatter correction (MSC),[19] standard normal variate (SNV),[20,21] Savitsky–Golay derivative[22] and their combinations were applied to the spectral pretreatment before PCA.

2.3.2. *Principal component analysis*

The classification of Nanfeng mandarin samples by geographical area requires a method that yields a positive identification, i.e., a sample should be classified as belonging to a class only if it is similar enough to that considered class.[23] In fact, the most commonly used methods for classification in chemometrics are the visual dimensional reduction methods based on latent projections. PCA[4,24] as a much-used such method for providing unsupervised visual classification was introduced in this study. It converts each NIR spectrum vector into a single point in principal component space (i.e., PCs), without losing the feature of data structure. Then, if the captured variance is relevant to chemical variations and sample classification, similar sample scores should cluster together on a graphical scores plot of PC 1 versus PC 2 or even PC 3.

2.3.3. *Changeable size moving window partial least squares*

NIR spectra often contain some irrelevant variables for classification, which may lessen both the accuracy and robustness of the models. Variable selection can discard signals that are not useful for classification, while primarily retaining signals that have information correlating with sample groups, to make the model simpler and obtain a better interpretation and lower measure system costs.[25,26] Up to now, there are many effective methods for variable selection.[27–32] CSMWPLS,[18] as one of them, is a strategy to search for an optimized subregion in spectral regions for producing better results. The superiorities of this method are: the window size is changeable and the window moves through the whole spectral region with fixed step. As shown in Fig. 1, the process of this algorithm is as follows: firstly, a spectral window that starts at the (i)th spectral channel and ends at the $(i + w - 1)$th spectral channel is constructed, where w is the window size. Then the window is moved

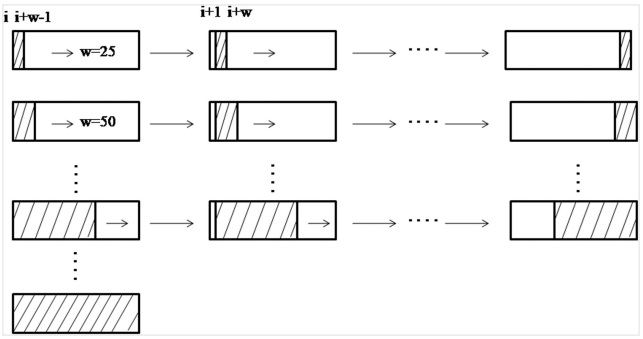

Fig. 1. The specific process of CSMWPLS algorithm.

through the whole region with a step of 1. Thus, there are $(n - w + 1)$ windows over the whole spectra, and with each window the calibration model is built with the corresponding subset of the spectral X. Afterward, the window size varies at an adjustable increment and the aforementioned procedures is run repeatedly with the new window size. After calculations for all the subsets, the region with the best prediction results is chosen as the informative region.

2.3.4. *Data processing of (CSMWPCA)*

According to the basic idea of CSMWPLS, the strategy of CSMW was utilized and replacing PLS with PCA to construct CSMWPCA. In this study, the window size varied from 20 to 800 with an interval of 5. The dataset of 583 samples was split as following: for each category, the samples were divided into calibration and validation sets by a ratio of ca. 9:1 (as shown in Table 1). Thus, in total 524 samples were taken as the calibration set, and the remaining 59 samples were selected to be the external validation.

When different pretreatment methods were applied to remove information not related to classification and CSMWPCA technique was used to select an optimized sub-region in spectral regions for producing better results, the quality of the PCA classification models were compared according to several evaluation parameters:

Total classification (prediction) rate:

$$TR = \frac{\sum_{i=1}^{k} m_{\mathrm{ci}}}{N} \tag{1}$$

Category rate:

$$RC_i = \frac{m_{\mathrm{ci}}}{N_{\mathrm{ci}}}. \tag{2}$$

These two equations were applied in both calibration and external validation sets, where m_{ci} and N_{ci} are the correct classification or prediction number and the total classification number of one category, respectively, and N is the total classification or prediction number of all categories. Meanwhile, graphical scores plots were also used to illustrate the goodness of the models.

Data pretreatments, variable selection and PCA in this study were carried out by self-editing programs in MATLAB (Ver. 7.1: The MATHWORKS, USA).

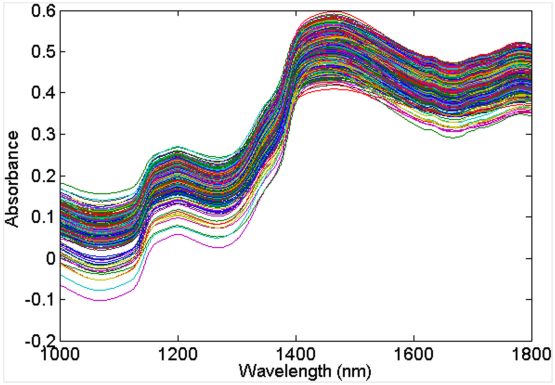

Fig. 2. Original NIR spectra of all Nanfeng mandorin samples.

3. Results and Discussion

3.1. *Near infrared spectra*

The original NIR spectra of all Nanfeng mandarin samples from seven different geographical areas are displayed in Fig. 2. All the achieved spectra data were averaged. No obvious differences were detected from a visual observation of the spectra among the seven category samples in the whole spectral range. And all samples have two significant absorption bands around 1190 and 1450 nm, which are generally assigned as the peaks of water, because of their high water content. Therefore, multivariate calibration techniques must be used for modeling based on near infrared spectra. And in this study, chemometric data reduction (CSMWPCA) and pattern recognition methods are a natural choice for analysis of such complex, inter-related NIR spectral data.

3.2. *PCA applied to raw data*

The PCA aimed to map the spectroscopy signals on to a low-dimension space with the largest variability. When PCA was applied to the raw NIR spectral data of the 524 mandarin samples in the calibration set, the scores plot was shown in Fig. 3. In this three-dimensional scores plot, the "coordinates", i.e., the scores on the first three principal components, provide a measurement of distance of each sample to all categories. In this way, each class model is defined by a class space delimited by the distance. Each model will accept samples whose distance to the corresponding central point is lower than that to other classes. Bearing in mind the categories studied here, as can be seen, class-models relating to all categories appear to be seriously overlapped, showing very poor separation.

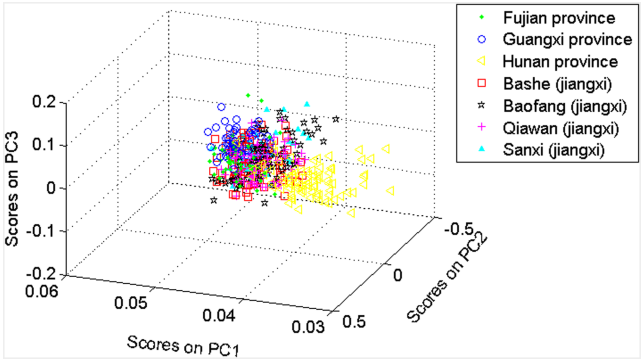

Fig. 3. Three-dimensional scores plot from PCA of raw NIR spectra for calibration samples.

These graphical results can also be confirmed numerically in Table 2. The total classification and prediction rate are 34.35% and 37.29%, and the category rates are from 6.74% to 74.07%. The fairly low correct rates reveal the problems for classification real samples in future, and serve for giving more sense to the aim pursued in this study, i.e., trying to improve the final classification model by wavelength selection to make possible a more accurate practical application.

3.3. *PCA after pretreatment and CSMWPCA variable selection*

In view of the results exhibited by the class models developed on the original NIR spectra, we decided to preprocess these spectra by different pretreatment methods and select informative wavelength as an attempt to improve classification performance. Table 2 summarizes the classification and prediction rates corresponding to the class models developed on the basis of raw NIR spectra and the spectra after diverse spectral pretreatment and variable selection methods. It can be seen in Table 2, comparing the results obtained from all the pretreatment methods, that the application of second derivative showed relative success to obtain a satisfactory classification model, providing 91.79% and 88.14% total correct classifications in both classification and prediction. Once second derivative was selected as pretreatment method to be used for correcting NIR spectra, CSMWPCA was applied to select useful wavelength. It can be seen in Table 2 that after the wavelength selection, the resulting class model showed an excellent discriminant power with 97.52% and 96.61% total classification and prediction rates and 100% category rates for C2, C4, C5, C6 and C7, which indicated a good clustering effect of mandarin samples from varied producing areas, and described the sample diversity by qualitative analysis.

These numerical results can be visually confirmed by scores plot (see Fig. 4) as well, showing a clear separation among classes and considerably improved with regard to the low model complexity (290 variables). All the mandarin samples are closely clustered in the each region of the PCs space, and the differences among most samples were pronounced. However, to samples from Fujian and Hunan province, i.e., C1 and C3, the diversity between them

Table 2. Percentages of correctly classified samples.

Pre-treatment	Wavelength region (nm)	Classification (%)								External prediction (%)
		RC1	RC2	RC3	RC4	RC5	RC6	RC7	TR	TR
Raw spectra	1000–1800	56.00	74.07	42.22	6.74	12.36	23.33	58.06	34.35	37.29
MSC	1000–1800	28.00	62.96	52.22	66.29	19.10	21.11	33.87	40.27	45.76
SNV	1000–1800	28.00	62.96	52.22	66.29	19.10	22.22	33.87	40.46	45.76
First derivative	1000–1800	56.00	92.59	66.67	11.24	17.98	28.89	38.71	40.84	45.76
Second derivative	1000–1800	80.00	98.15	64.44	100.00	100.00	100.00	100.00	91.79	88.14
MSC + first derivative	1000–1800	62.00	90.74	70.00	11.24	22.47	28.89	37.10	42.37	47.46
MSC + second derivative	1000–1800	56.00	100.00	64.44	87.64	93.26	92.22	85.48	83.40	79.66
SNV+ first derivative	1000–1800	62.00	90.74	70.00	11.24	22.47	28.89	37.10	42.37	47.46
SNV+ second derivative	1000–1800	56.00	100.00	64.44	87.64	93.26	92.22	85.48	83.40	79.66
Second derivative	1007–1296	92.00	100.00	90.00	100.00	100.00	100.00	100.00	97.52	96.61

Note: Total rate (TR) in classification and external prediction, category rates (RC1, RC2, RC3, RC4, RC5, RC6, RC7) in classification.

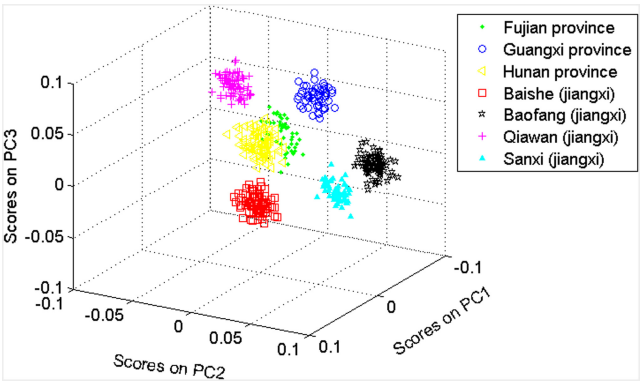

Fig. 4. Three-dimensional scores plot from PCA of NIR spectra after optimal pretreatment and CSMVPCA variable selection for calibration samples.

was not such notable as others. Some samples are slightly overlapping. Nevertheless, in short, the PC1, PC2 and PC3 provide a good image for the geographical classification of Nanfeng mandarins.

4. Conclusion

This study has illustrated the feasibility of applying NIRS combined with PCA to classify Nanfeng mandarin samples by geographical region. The success of this strategy depends largely on wavelength selection method which not only significantly enhances the quality of the classification model in terms of accuracy, but also makes the model simpler. The classification model constructed based on the second derivative pretreated spectra and CSMWPCA selected wavelength region prompted a substantial improvement in comparison with the model developed based on the original spectra. The notably improved model showed total classification rate of 97.52%, and a good prediction ability of 59 samples in an independent test set with the total prediction rate of 96.61%.

The promising results reported in this study may serve to support the feasibility of providing a straightforward, fast and objective Nanfeng mandarin authentication and determination of producing areas of unknown samples.

Acknowledgment

This work was supported by General Administration of Quality Supervision, Inspection and Quarantine of the People's Republic of China (2012IK169) and National Natural Science Youth Foundation of China (21205053).

References

1. S. Tsuchikawa, "Near-Infrared Spectroscopy in Food Science and Technology," Y. Ozaki, W. F. McClure, A. A. Christy, Eds., pp. 138–139, John Wiley & Sons, Inc, Hoboken, New Jersey, U.S. (2006).
2. D. A. Burns, E. W. Ciurczak, "Handbook of Near-Infrared Analysis," Taylor & Francis Group, Boca Raton, U.S. (2008).
3. T. Caliński, J. Harabasz, "A dendrite method for cluster analysis," *Commun. Stat.-Theor. M* **3**(1), 1–27 (1974).
4. S. Wold, K. Esbensen, P. Geladi, "Principal component analysis," *Chemometr. Intell. Lab.* **2**(1–3), 37–52 (1987).
5. K.-L. Hsu, H. V. Gupta, S. Sorooshian, "Artificial neural network modeling of the rainfall-runoff process," *Water Resour. Res.* **31**(10), 2517–2530 (1995).
6. I. Estebandiez, "An evaluation of orthogonal signal correction methods for the characterisation of arabica and robusta coffee varieties by NIRS," *Anal. Chim. Acta.* **514**(1), 57–67 (2004).
7. I. Esteban-Diez, J. M. Gonzalez-Saiz, C. Saenz-Gonzalez, C. Pizarro, "Coffee varietal differentiation based on near infrared spectroscopy," *Talanta* **71** (1), 221–229 (2007).
8. D. Wu, Y. He, Y. Bao, "Intelligent Computing," pp. 931–936, Springer (2006).
9. Y. He, X. Li, X. Deng, "Discrimination of varieties of tea using near infrared spectroscopy by principal component analysis and BP model," *J. Food Eng.* **79**(4), 1238–1242 (2007).
10. Y. Shao, Y. He, Y. Wang, "A new approach to discriminate varieties of tobacco using vis/near infrared spectra," *Eur. Food Res. Technol.* **224**(5), 591–596 (2007).
11. T. Woodcock, G. Downey, J. D. Kelly, C. O'Donnell, "Geographical classification of honey samples by near-infrared spectroscopy: A feasibility study," *J. Agric. Food. Chem.* **55**(22), 9128–9134 (2007).
12. G. Downey, P. McIntyre, A. N. Davies, "Geographic classification of extra virgin olive oils from the eastern Mediterranean by chemometric analysis of visible and near-infrared spectroscopic data," *Appl. Spectosc.* **57**(2), 158–163 (2003).
13. G. Gurdeniz, B. Ozen, "Detection of adulteration of extra-virgin olive oil by chemometric analysis of mid-infrared spectral data," *Food Chem.* **116**(2), 519–525 (2009).

14. L. Liu, D. Cozzolino, W. Cynkar, M. Gishen, C. B. Colby, "Geographic classification of Spanish and Australian Tempranillo red wines by visible and near-infrared spectroscopy combined with multivariate analysis," *J. Agric. Food Chem.* **54**(18), 6754–6759 (2006).
15. L. Pillonel, W. Luginbühl, D. Picque, E. Schaller, R. Tabacchi, J. Bosset, "Analytical methods for the determination of the geographic origin of Emmental cheese: Mid and near-infrared spectroscopy," *Eur. Food Res. Technol.* **216**(2), 174–178 (2003).
16. L. León, J. D. Kelly, G. Downey, "Detection of apple juice adulteration using near-infrared transflectance spectroscopy," *Appl. Spectosc.* **59**(5), 593–599 (2005).
17. Y. L. Wei, C. H. Yin, G. P. Chen, J. Huang, W. B. Zhang, Y. P. Du, "Identification of Nanfeng Mandarin from different origins by using near infrared spectroscopy coupled with principal components analysis," *Spectrosc. Spect. Anal.* **33**(11), 3024–3027 (2013).
18. Y. Du, Y. Liang, J. Jiang, R. Berry, Y. Ozaki, "Spectral regions selection to improve prediction ability of PLS models by changeable size moving window partial least squares and searching combination moving window partial least squares," *Anal. Chim. Acta* **501**(2), 183–191 (2004).
19. T. Isaksson, T. Næs, "The effect of multiplicative scatter correction (MSC) and linearity improvement in NIR Spectroscopy," *Appl. Spectosc.* **42**(7), 1273–1284 (1988).
20. I. A. Cowe, J. W. McNicol, D. C. Cuthbertson, "A designed experiment for the examination of techniques used in the analysis of near infrared spectra. Part 1 Analysis of spectral structure," *Analyst* **110** (10), 1227–1232 (1985).
21. R. Barnes, M. Dhanoa, S. J. Lister, "Standard normal variate transformation and de-trending of near-infrared diffuse reflectance spectra," *Appl. Spectosc.* **43**(5), 772–777 (1989).
22. A. Savitzky, M. J. Golay, "Smoothing and differentiation of data by simplified least squares procedures," *Anal. Chem.* **36**(8), 1627–1639 (1964).
23. D. Hand, "Discrimination and classification," John Wiley & Sons, Toronto (1981).
24. H. Abdi, L. J. Williams, "Principal component analysis," *WIREs Comput. Stat.* **2**(4), 433–459 (2010).
25. C. M. Andersen, R. Bro, "Variable selection in regression—a tutorial," *J. Chemometr.* **24**(11–12), 728–737 (2010).
26. C. H. Spiegelman, M. J. McShane, M. J. Goetz, M. Motamedi, Q. L. Yue, G. L. Coté, "Theoretical justification of wavelength selection in PLS calibration: Development of a new algorithm," *Anal. Chem.* **70**(1), 35–44 (1998).
27. A. S. Bangalore, R. E. Shaffer, G. W. Small, M. A. Arnold, "Genetic algorithm-based method for selecting wavelengths and model size for use with partial least-squares regression: Application to near-infrared spectroscopy," *Anal. Chem.* **68**(23), 4200–4212 (1996).
28. L. Norgaard, A. Saudland, J. Wagner, J. P. Nielsen, L. Munck, S. Engelsen, "Interval partial least-squares regression (iPLS): A comparative chemometric study with an example from near-infrared spectroscopy," *Appl. Spectosc.* **54**(3), 413–419 (2000).
29. W. Cai, Y. Li, X. Shao, "A variable selection method based on uninformative variable elimination for multivariate calibration of near-infrared spectra," *Chemometr. Intell. Lab.* **90**(2), 188–194 (2008).
30. H. Li, Y. Liang, Q. Xu, D. Cao, "Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration," *Anal. Chim. Acta.* **648**(1), 77–84 (2009).
31. K. Zheng, Q. Li, J. Wang, J. Geng, P. Cao, T. Sui, X. Wang, Y. Du, "Stability competitive adaptive reweighted sampling (SCARS) and its applications to multivariate calibration of NIR spectra," *Chemometr. Intell. Lab.* **112**(0), 48–54 (2012).
32. H. Xu, Z. Liu, W. Cai, X. Shao, "A wavelength selection method based on randomization test for near-infrared spectral analysis," *Chemometr. Intell. Lab.* **97**(2), 189–193 (2009).